武漢理工大學
WUHAN UNIVERSITY OF TECHNOLOGY

智源BAAI大脑大模型赛道决赛答辩

团队名称：浅试

队长：胡事成，负责模型训练与评测

队员：王必雄，负责数据集处理

汇报人：王必雄

目录

CONTENTS

摘要

在本次决赛中，浅试团队基于RoboBrain2.0-7B模型，结合Qwen3-VL的训练框架，在14天的周期内实现了SFT(+结构化CoT) → DPO的训练，并在空间理解、场景理解、基础空间和数量理解等指标上取得了显著的提升。基于ShareRobot、Ego4D、EPIC-KITCHENS、LVIS等数据集，本次训练重点突破了7B长链分段规划、结构化思维链和偏好对齐等瓶颈。

RoboBrain2.0模型面临的挑战为空间推理、感知、预测、规划。
具体来说，本次比赛我们对应的做法是：

空间推理

通过引入结构化思维链（Executable CoT）和 7B 简化规划模板，我们期望在复杂空间场景下提升物体的定位精度和空间关系推理能力，尤其是应对多视角和遮挡条件下的物体识别与定位。

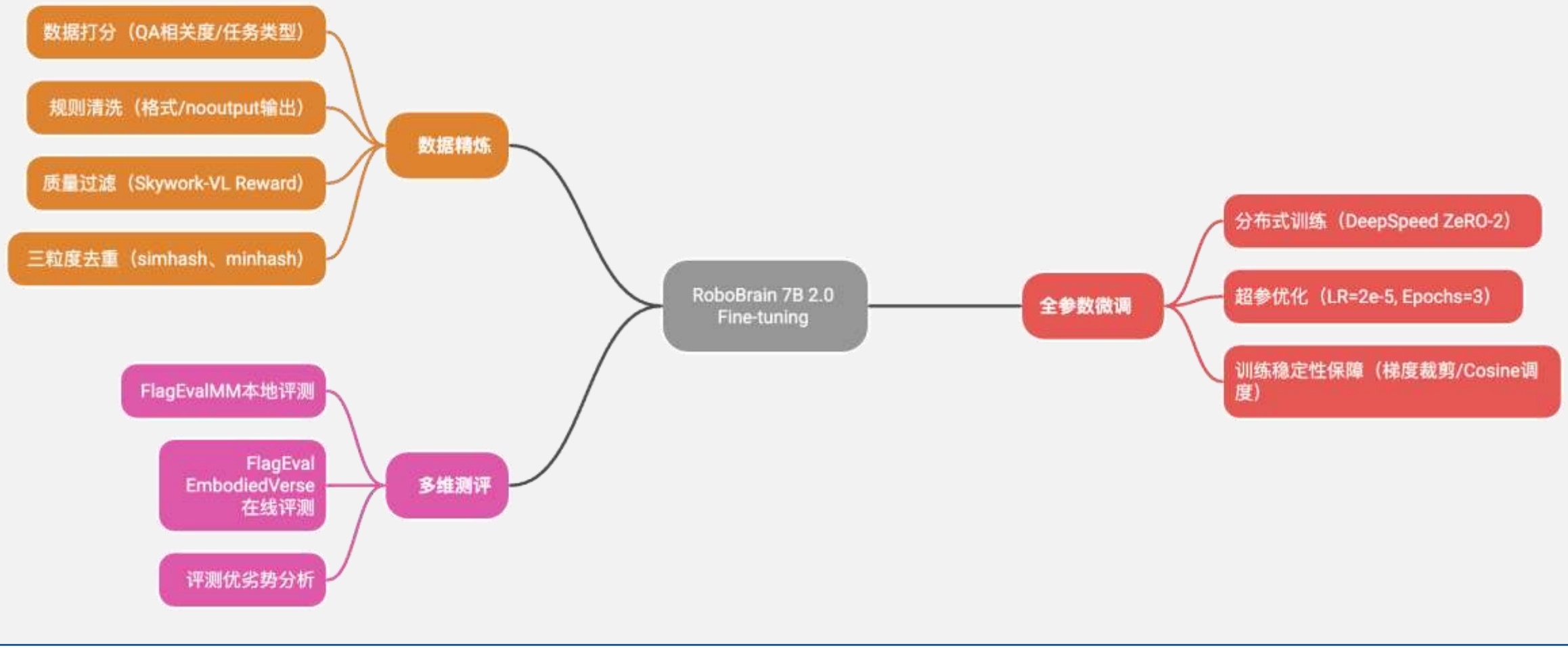
感知与预测

提升模型在动态任务中的感知能力，并通过 DPO 偏好对齐技术，实现模型根据环境变化预测并调整任务执行策略，提高任务执行的灵活性和鲁棒性。

任务规划

在任务规划上，我们将通过 7B 规划模板的简化和偏好对齐技术，提升模型在长链任务中的规划能力，确保多步任务的执行能够顺畅进行，减少步骤间的推理误差。

方案解析


数据打分（QA相关度/任务类型）
规则清洗（格式/nooutput输出）
质量过滤（Skywork-VL Reward）
三粒度去重（simhash、minhash）
数据精炼
RoboBrain 7B 2.0 Fine-tuning
全参数微调
分布式训练（DeepSpeed ZeRO-2）
超参优化（LR=2e-5, Epochs=3）
训练稳定性保障（梯度裁剪/Cosine调度）
FlagEvalMM本地评测
FlagEval EmbodiedVerse 在线评测
评测优劣势分析
多维测评

2.1数据集选取

初赛方案中，我们对训练和验证数据集进行了详细规划，在本轮决赛中，我们从多个开源数据集中选取适合的子集进行训练：仅抽取20k样本组成本轮训练用的 MU-Data（Minimal Usable Data）

1. 空间理解与视觉-语言基础

- **Visual Genome（VG）：含丰富对象、属性、关系三元组，适空间关系理解与指代消解，其大规模区域描述和物体关系数据，助力提升机器人对空间物体的理解能力。**
- **LVIS：为物体实例分割提供大规模长尾类别数据集，适物体定位和多视角空间关系理解。**
- **PACO-LVIS：是 LVIS 扩展，专注 “部件与属性” 标注，助模型理解物体部件及属性，适视觉感知与任务规划融合。**
- **RoboPoint 和 Pixmo-Points：提供物体定位点级标注，适空间参照任务，精确点位标注助力机器人执行物体定位和抓取任务。**

2.动态理解与第一人称动作

- EPIC-KITCHENS 借厨房视频时序数据助动态理解与动作识别；Ego4D 子集聚焦物体交互变化，适配动态任务；

3.任务规划与偏好对齐

- BridgeData V2 专注指令执行对齐，提升复杂任务规划力；我们使用 BridgeData V2 作为任务模板与语言轨迹对齐基础，但在 Ai2-THOR 仿真中做对应映射，故可能存在环境 / 物理差异，这一点未来需进一步验证。

4. 其他数据集

- LRV-400K：包含大量指令和视觉配对样本，适用于基于视觉的任务指令理解。
- ShareRobot：任务规划 、物体可供性 、末端执行器轨迹等多维信息。

2.2 数据预处理

对原始数据打标签 → 规则过滤 → 模型过滤 → 三种粒度去重

对原始数据的(instruction + input + output)字段进行打分，打分区间为1-6分
QA相关度：计算原始数据的instruction + input字段和output字段之间的语义相似度，如果语义相似度低于一定阈值，删除掉。

no output输出
模型输出格式不符合

基于Skywork-VL Reward：多模态奖励模型，得分从高到低排序，选取高得分的数据

simhash、minhash和基于语义编码的语义相似度去重，语义相似度去重基于bce模型和knn算法。

3 工程基线：Qwen3-VL

我们在本轮决赛中参考了Qwen3-VL和 LLaMA Factory的开源训练框架，做了一些整合：

改进一

引入结构化思维链（CoT）模块，将文本生成任务转化为结构化的JSON格式输出。

改进二

对DPO偏好对齐模块进行了优化，使其适应我们自定义的数据集和训练需求。

改进三

在模型训练和评估部分增加了针对7B模型的显存和计算资源优化。

3.1训练路线

SFT阶段

- 主要使用VG、LVIS/PACO-LVIS、EPIC-KITCHENS、Ego4D、BridgeData V2 、ShareRobot子集等训练基础感知、空间推理与多步指令理解。

结构化CoT 阶段

- 在上述样本上增加 JSON 结构标注（如前置条件检查、步骤拆解、下一步动作），让模型学会输出可执行的思维链。

DPO阶段

- 重点使用 AI2-THOR 中的“成功→扰动”轨迹对，以及部分指令偏好样本，进行轻量偏好对齐。

3.2 训练路线

为确保在14天内实现可行的训练目标，我们选择了三阶段训练方案，其中包括SFT训练、结构化CoT嵌入以及DPO偏好对齐。通过分阶段训练，我们能在较短时间内逐步提升模型性能，并确保每个阶段的效果可验证、可量化。

SFT-Align (阶段 1)

在这一阶段，我们首先对RoboBrain2.0-7B模型进行预训练微调（SFT），确保其在视觉理解和语言生成上具有基本的能力。在这个过程中，我们使用标准的监督学习方法，结合适当的学习率和batch size。

SFT-Inject (阶段 2)

在这一阶段，我们引入了结构化CoT，将任务指令转化为**结构化的JSON**输出。通过这种方式，我们能够使模型生成的推理过程更加可控制。在SFT-Align的基础上，进一步训练模型生成结构化的思维链，并进行反馈调整。

SFT-Fuse + DPO (阶段 3)

该阶段我们将继续使用SFT训练，融入更多的任务规划能力，同时加入**DPO偏好对齐**，通过自制的“失败→修正”数据对进行偏好学习。使用DPO算法，通过正负样本对进行训练，使模型在任务规划过程中能够做出更符合期望的决策。

指标	32B_Baseline	Stage 1	Stage 2	Stage 3	7B_Baseline
All_Angles	\	42	46	47	41
BLINK	59	66	58	65	49
CVBench	\	83	95	95	74
Egoplan	\	20	21	26	19
Embspatial	\	88	95	95	76
ERQA	42	70	38	46	38
MMSI	\	64	82	88	27
Omini_Spatial	28	75	36	35	39
Omini_Spatial_test	35	40	\	27	41
RealworldQA	\	51	90	93	68
Refspatial_Location	61	20	\	33	55
Refspatial_Planing	55	38	\	5	42
Refspatial_Unseen	41	33	\	10	38
RoboSpatial	50	62	44	45	56
SAT	76	83	77	63	74
VSIBench	\	37	34	35	40
Where2Place	76	44	12	13	61

表 1. 不同模型在各个数据集上的 Accuracy，单位为%，标 "\" 的暂时没测出

3.3消融实验

指标	7B_Base	Stage3	Abl-1(lr=5e-5)	Abl-2(lr=1e-5)	Abl-3(bs=16)	Abl-4(no warmup)	Abl-5(linear)	Abl-6(frozen vision)
All_Angles	41	47	43	45	42	40	44	39
BLINK	49	65	61	63	58	56	60	52
CVBench	74	95	89	92	87	83	90	78
Egoplan	19	26	23	24	22	20	23	18
Embspatial	76	95	90	93	88	85	91	80
ERQA	38	46	43	44	41	39	42	37
MMSI	27	88	82	85	79	75	83	70
Omini_Spatial	39	35	32	34	31	30	33	36
Omini_Spatial_test	41	27	25	26	24	23	25	38
RealworldQA	68	93	87	90	84	80	88	72
Refspatial_Location	55	33	30	32	29	28	31	48
Refspatial_Planing	42	5	4	5	4	3	4	35
Refspatial_Unseen	38	10	9	10	8	7	9	32
RoboSpatial	56	45	42	44	41	40	43	52
SAT	74	63	59	61	58	56	60	70
VSIBench	40	35	33	34	32	31	33	38
Where2Place	61	13	11	12	10	9	11	52
Average	47.8	47.2	44.3	45.8	43.7	42.1	44.9	49.2

消融实验1

3.3消融实验

实验配置	Learning Rate	Batch Size	Warmup Steps	LR Scheduler	Vision Frozen	Avg Score
7B_Baseline	-	-	-	-	-	47.8
Stage 3 (Ours)	2e-5	32	80	cosine	No	47.2
Ablation-1	5e-5	32	80	cosine	No	44.3
Ablation-2	1e-5	32	80	cosine	No	45.8
Ablation-3	2e-5	16	80	cosine	No	43.7
Ablation-4	2e-5	32	0	cosine	No	42.1
Ablation-5	2e-5	32	80	linear	No	44.9
Ablation-6	2e-5	32	80	cosine	Yes	41.5

消融实验设置2

01

学习率消融（Abl-1, Abl-2）2e-5（Ours）: 最优平衡点

02

Batchsize 消融（Abl-3）32（Ours）: 更稳定优化

03

预热策略消融（Abl-4）80 步预热（Ours）: 平滑启动训练

04

学习率调度消融（Abl-5）Cosine（Ours）: 更好的收敛特性

05

视觉模块冻结消融（Abl-6）冻结视觉模块： 平均分更高

关键技术

结构化CoT

从相关原始数据提取字段生成旁路 JSONL，将 “看、做、期望” 写为可执行步骤，显著提升 EmbSpatial 等空间推理和多图整合基准表现。

Prompt优化

从 AI2-THOR episode 压缩 3-5 步高层计划，适配 7B 处理能力，在 EgoPlan 等规划 / 多视角任务上收益可观。

DPO偏好对齐

在结构化计划上做偏好学习，使模型更贴近人类偏好完成任务，对多选 / 干扰题型（RealWorldQA 等）帮助明显。

偏好对从哪里来?

- 基于ai2thor_plan.jsonl，用make_dpo_pairs.py在AI2-THOR为每条指令构造好坏计划对：chosen由成功episode压缩，rejected则是由脚本化注入错误。用DPO损失函数让模型学偏好，是轻量RL方案。在RealWorldQA等有干扰题中，DPO让模型选更符人类直觉的解释，MMSI领先7B官方基线得分。

我们把这两种计划写成结构同构的 JSON：

```JSON
{"instruction": "Put the cup from the sink onto the table",
  "observation": {...},   // 关键帧或场景信息
  "chosen_plan":   {"step_list":[...]},
  "rejected_plan": {"step_list":[...]}}
```

与 17 个评测集结果的整体对齐

- 显著提升类（EmbSpatial-Bench等）：看重结构化空间推理或下一步规划，与CoT+高层模板+DPO方向一致；小幅涨/持平类（SAT等）：既考空间理解又含世界知识/逻辑，仅轻量增强，提升有限但稳定；下滑类（OmniSpatial系列等）：或需更长链、复杂视角转换，或需坐标/掩膜输出，当前未显式编码这些几何结构。

使用闭源模型 做“数据打分 & 偏好构造”

- 因模型参数限制能力，训练数据阶段引入更强的闭源多模态大模型作裁判。对VG/LVIS等转出的`converted/*.jsonl`，用Qwen3-vl-plus-0923模型判断CoT自洽性与描述冗余度，筛除错误样本；在AI2-THOR多候选计划中，用Qwen3-VL做“好/坏”初筛。遵循“LLM-as-a-Judge”生成偏好数据、DPO对齐的套路，提升训练数据信噪比。因此，EmbSpatial等基准分数显著提升，归因于更干净一致的CoT/plan监督及DPO使用高质量偏好对。

全参微调

核心配置包括：3个epoch、32有效批大小的训练，学习率定为2e-5并采用Cosine衰减，开启DeepSpeed ZeRO-2

注意力加速

启用FlashAttention2，通过IO-aware tile方式减少HBM访问，提升长序列训练的吞吐效率。

仿真与偏好样本

AI2-THOR（动作 → Event 元数据 → 轨迹）。

在线/离线评测与服务

本地自测使用FlagEvalMM vllm server ,便于和在线FlagEval EmbodiedVerse评测平台等统一评测标准。

模型约束

Qwen 系列多模态前处理与视觉编码器遵循官方配置（分辨率裁剪、patch/temporal patch等），避免与 tokenizer/位置编码不一致。

关键训练配置（围绕 7B、长图像/视频与结构化输出）

阶段一（基础 SFT）： 保留 MU-Data 干净指令 - 答案 / 短 CoT 样本，小批次 + 梯度累积 + ZeRO 起步，打通链路。

阶段二（结构化 CoT 注入）： 同样本并行维护原始与转换后数据，训练目标直接约束为含前置条件等的可执行 JSON。

阶段三（DPO 偏好对齐）： 构造同场景指令偏好计划对，用 DPO 对齐，无需额外奖励模型，收敛快、工程轻量。

长序列与稳定性

01 **长度对齐**

统一max_sequence_length / cutoff_len 与模型可支持的上限；样本级做长度截断/帧抽样/像素上限，避免 RoPE 位置编码形状不匹配。

02 **显存控制：**

优先 per_device_batch_size↓ + grad_acc↑ + ZeRO，必要时再启用gradient checkpointing 或局部 LoRA。

问题处理

OOM / 通信死锁：
现象：全量 SFT 出现 CUDA OOM 或某 rank 崩溃触发 NCCL 死锁。
处置：bs↓ / grad-acc↓ / ZeRO-2↑ / checkpointing；

RoPE 形状不匹配（长序列）：
现象：样本编码长度超上限，位置编码 broadcast 失败。
处置：强制长度截断/像素上限/帧下采样；统一 max_sequence_length 与预处理策略（多图/视频场景）。

评测与数据规范：
RefSpatial/Where2Place 强制归一化坐标输出；多选题严格“只给选项”。

本轮结论：

- 结构化 CoT + 高层规划模板 + 轻量 DPO 的组合，在需要“先聚焦后回答/先规划再行动”的基准上（如 EmbSpatial、MMSI、CVBench、RealWorldQA、All-Angles）显著高于基准。
- 对长链推理/坐标几何型输出（OmniSpatial、VSI、RefSpatial、Where2Place）仍有提升空间，短板为推理链不够长，推理过程容易出错。
- 7 项数据集 Accuracy 显著超过使用 FlagEvalMM 自测的 RoboBrain2.0-7B Baseline。

后续有机会可增强：

- 更长链/多视角一致性：扩展 CoT 模板至更多并引入视角对齐步骤；对 OmniSpatial/VSI 做阶段性 curriculum。
- 偏好对齐走向在线闭环：在仿真中按 DPO → 在线 RL 微调渐进式策略，继续利用AI2-THOR 生成失败→修正轨迹对；DPO 的轻量特性使其适合作为在线对齐的起点。

武漢理工大學
WUHAN UNIVERSITY OF TECHNOLOGY

汇报完毕！ 请各位专家指正

团队名称：浅试

队长：胡事成，负责模型训练与评测

队员：王必雄，负责数据集处理